Data Capture

Guidelines for Capture of RSC Journal Articles

Also see: www.rsc.org/authorguidelines

CONTENTS

1.0 Introduction

1.1 Scope of this document

These guidelines apply to Version 3.8 of the RSC Article document type.

1.2 Feedback and updates

Please let us know of any problems you encounter in using these instructions while trying to encode articles using the schema provided. This will help us to improve both the application and its associated documentation.

$$\int_{j}^{i} xd(\sqrt[3]{?})$$

We plan to issue updates to the schema and documentation at regular, planned intervals. You will be notified of these updates in advance, so that you can allocate resources to deal with any changes to data capture instructions or rendering software that might be required.

1.3 Format of this document

This document fulfils two functions. As well as containing instructions on the conventions to follow, it acts as an example, being written to conform to the RSC Primary Articles document type Version 3.8.†

2.0 Scope of the data capture work

The objective is to capture all the text in each article in XML (see next section). The document element will always be `<article>`. Within this, the `<art-admin>` (for the article's unique manuscript number), `<published>` (for articles which have already been published), `<art-front>`, `<art-body>` and `<art-back>` elements will be routinely used, with an occasional `<appmat>`.

2.1 XML encoding

As far as possible, all the information in the articles presented should be encoded in XML and included in the resulting document. Obvious exceptions are figures, which should be referenced as external graphic files (see Graphics below).

Both tables and equations are liable to be more difficult. If possible, these should be encoded in XML, but we accept that there are liable to be cases where this is not possible due to the complexity of the data or inadequacies in the schema as currently drafted. In these cases the relevant object should be treated as a graphic. A particular example is where a table contains graphics spanned across rows or columns - this would be impossible to render accurately from the XML. See Tables and Equations below for specific guidelines.

Articles should conform to XML conventions. This means that:

- an XML declaration must be provided at the start of the article;
- processing instructions must be terminated by '`?>`';

† The XML version of this document can be viewed using a suitable browser at http://www.rsc.org/schema/rscart38/captureguidelines.xml

- empty elements must be terminated by '/>' (*i.e.* for <colspec>, <ugraphic>, <icgraphic>);
- end tags should always be provided, except for empty elements;
- attribute values should always be quoted.

A variety of tools can (and should) be used to check that articles consist of valid XML. For example, if you open an XML document in Internet Explorer 5 or above, its built-in XML parser will check the document for validity and report any errors.

2.2 File naming conventions

All manuscripts will have a unique identifier, assigned by RSC, *e.g.* b701234h. As well as being used to name the file containing the encoded article, this identifier will be encoded as the <ms-id> element within the article.

The RSC will name graphics files as follows:

- b701234h-f1.tif (for figure 1);
- b701234h-s2.tif (for scheme 2);
- b701234h-u1.tif (for ugraphic 1);
- b701234h-t1.tif (for ugraphic 1 created by supplier).

Graphic types (from RSC):

- f1, 2, 3 ... figures;
- s1, 2, 3 ... schemes;
- u1, 2, 3 ... ugraphics;
- t1, 2, 3 ... ugraphics created by supplier;
- ga graphical abstract;
- c1, 2, 3 ... charts;
- p1, 2, 3 ... plates.

The following naming styles should be supplied to the RSC:

- for <ms-id> use the form 'b701234h';
- for <doi> use the form '10.1039/b701234h';
- for the XML files (and PDF) use the filenames in the form 'b701234h.xml', 'b701234h.pdf'
- for graphics generated at data capture (maths, table images where required) use the form 'b701234h-t1.tif', and increment the numbers as t2, t3, t4, *etc.* through the document.

Lower-case should be used for all of the above.

2.3 File delivery

We require, for each paper:

- An XML file named as b701234h.xml (file width - max width 1000 characters);
- Mathtype files created for inline/displayed maths, as GIF files at 600 dpi. The XML file will require call-outs to these images. The images should be named as specified above (*e.g.* b701234h-t1.tif).

Each document and associated files should be delivered as a zip file, named as above (*e.g.* b701234h.zip)

3.0 Documents relating to the RSC document type

The schema itself is in the file rscart38.xsd. A number of other files are required before documents will parse against the schema. Two of them (calstab1.xsd and class.xsd) should be stored in the same directory as the schema itself, while the folder containing the MathML schema (mathml2) should be stored at the same directory-depth as the folder for the RSC schema.

3.1 Table support

The file calstab1.xsd contains the OASIS-supported schema fragment which supports the interoperable CALS table model subset. Additions and changes to this model are declared in the body of the schema itself, not here.

3.2 Special characters

A file containing allowable character entities is provided; this contains mappings of characters to numeric values that conform to Unicode 2.0 (rsc_x.ent). Entities based on the Unicode value should be used (*e.g.* `&#x3b1;`) in the XML. It should be noted that we use Unicode combining characters to partially solve the problem of 'one character over another'. This means that rendering software will need to support combining characters, ideally in a generalized manner.

If an article contains any characters or symbols which are not in the RSC set, the relevant editorial team should be contacted. They will be able to advise on a suitable course of action, and will set in motion their addition to the standard set.

4.0 General conventions

4.1 Guidelines

Semantics of the table model. The table model used is developed from the interoperable CALS table model subset supported by OASIS.[1,2] The OASIS web site contains a description of the generic CALS table model,[1] and a description of the semantics of this interoperable subset.[2]

The document type simplifies the level of CALS table support that is required by removing the <spanspec> element type (which is not part of the interoperable subset). This has been found to be unnecessary, since both horizontal and vertical spans within tables can be represented without it. (<colspec> provides all the information that is required for horizontal spanning, while the morerows attribute supports vertical spans.) It adds support for rotated tables by including the orient attribute, which can be set to "land" to indicate a landscape, *i.e.* rotated, table.

4.2 Article structure

Each article consists of front matter, body matter and back matter.

The article itself can have a type attribute, which specifies what type of article it is. Table 1 summarises the codes to be used for each type of article.

4.3 Front matter

The front matter consists of <art-admin>, which holds the article's unique manuscript number, <published>, which contains details of the journal, volume, issue in which the article has been printed and the relevant pagination details, and <art-front>, which is the front matter proper.

For accepted date, use:

```
<date role="accepted">
  <year>2007</year>
  <month>April</month>
  <day>23</day>
</date>
```

For the date on which a revised version of an article was issued, use the same date format, with role="revised".

We would like the corresponding author to be identified, using role="corres".

For affiliations, use:

```
<org><orgname></orgname></org>
<address></address>
```

for the address—although the <org> group does contain its own <address> element, this shouldn't be used for encoding the articles.

The <published> element should be set with the attribute type="print", along with the journal code—the other subelements should be left blank, thus:

```
<published type="print">
  <journalref><link>GC</link></journalref>
  <volumeref><link></link></volumeref>
  <issueref><link></link></issueref>
  <pubfront>
    <fpage></fpage>
    <no-of-pages></no-of-pages>
    <date><year></year></date>
  </pubfront>
</published>
```

4.4 Body matter

The body of each article consists of an <art-body>, containing one more <section> elements.

Table 1 Article types and their corresponding codes

Article type	Code
Article	ART
Communication	COM
Perspective	PER
Letter	LET
Feature Article	FEA
Editorial	EDI
Obituary	OBT
Interface	INF
Method	MET
Frontiers	FRO
Discussion	DIS
Review Article	REV
Emerging Area	EMA
News Item	NWS
Highlight	HIG
Application	APP
Metals in Perspective	MES
Interview	INT
Technical Note	TEC
Conference Diary	CNF
Conference Report	CRP
Synthetic Abstract	SAB
Cover Feature	COV
Profile	PRO
Focus	FOC
Viewpoint	VPT
Keynote Article	KEY
Hot of the Press	HOT
ASU Article	ASU
Analytical Methods Committee	AMS
Critical Review	CRV
Tutorial Review	TRV
Comment	CMT
Invited Article	INV
Opinion	OPN
Education	EDU
Forum	FOR
40th Anniversary Article	AFO
Essay	ESS
Air in Perspective	AIR
Water in Perspective	WAT
Book Chapter	BOK

These are the top-level structural units within each article: lower levels are represented by <subsect1>, <subsect2>, *etc.* (**N.B.** the numbering of section-level element names represents their depth of nesting, not repetition.)

Care should be taken to ensure that the structure of the article, implied by the style of headings, is correctly reflected in the <section> and <subsect*x*> elements assigned. See <title> for details of heading typestyles.

4.5 Appendices

Any appendices are placed within an <appmat> element, between the <art-body> and <art-back> elements. This contains one or more <appendix> elements, each optionally numbered and containing one or more <section> elements.

4.6 Back matter

The back matter contains an optional <ack> element, followed by mandatory <biblist>, <compoundgrp>, <annotationgrp>, <datagrp> and <resourcegrp> elements.

The <compoundgrp> element is provided as a place to collect together <compound> elements, each of which defines the ID of a chemical compound mentioned in the article, and thus to provide a target for <compoundref> cross-references (which are normally set in bold face: see Cross-references).

4.7 Graphics

Graphics take the following attributes:

- id: a unique ID for this graphic (see notes below on assigning IDs) (required)
- xsrc: the filename of the associated graphic

Other attributes are available but shouldn't be used.

Chemical formulae, equations and symbols for which no character entity is provided in the schema and tables which are too complex to encode as XML should all be encoded as a <ugraphic> element. As well as the standard attributes for graphics, this has a display attribute. If the <ugraphic> forms part of the current line of text, this should be set to "inline"; if the graphic needs be set off from the surrounding text, the attribute should be left blank (a value of "displayed" is assumed when the attribute is empty).

4.8 Assigning unique ids

In order to make ids unique within each article, a system of standardised prefixes and integers should be used.

4.9 Links and cross-references

Internal cross-references within an article should use the standard XML ID-IDREF mechanism. To enforce this, we have specified as required the id attribute for all the elements that cross-references might point to. Although the idrefs attribute is not mandatory in the corresponding reference elements, it should always be filled.

This table summarises the element types which indicate cross-references, and the target element type for each.

One specific point to note is that <citref> does *not* point to a <citation> or <journalcit> element: instead it points to <citgroup>. This design allows any number of citations to occur within a single numbered or sub-numbered part of a References list.

In the event that an DOI reference to another article is available, the general-purpose <link> element type is provided. In these cases the type attribute should be set to "doi".

Recognising cross-references. This table summarises typographical conventions which are often used to represent various types of cross-reference. Where a change of font style indicates such a cross-reference, it should always be marked up as such. In such cases, the cross-reference should *not* also be marked up as a change of font style.

4.10 Numbering

Numbers should be included in the <no> element if they are required.

There is no need (and no opportunity!) to number figures, schemes, boxes or plates. Suitable prefixes and numbers (*e.g.* "Fig 1") will be supplied by style sheets. Other concepts (*e.g.* citations, equations, appendices, and chemical compounds) have an optional <no> element. This does not need to be used where the numbering scheme follows a simple sequence of arabic numbers, since the entries will be auto-numbered in this case. If **any** instance of a given element type has a non-standard number within an article, then the <no> element should be specified for **all** instances of that element type.

Table 2 id prefixes for different classes of target

Type	Format[a]
author affiliation	affy
chart	chtx
chemical compound	chemx
citation	citx(y)[b]
equation	eqnx
figure	figx
footnote	fnx
plate	plx
scheme	schx
table	tabx
table footnote	tabxfny[c]
untitled graphic	ugrx
typesetter-generated graphic (*e.g.* equations and tables which cannot be encoded in XML)	ugty

[a] *x* denotes a sequence of integers (1, 2, 3 *etc.*), while *y* denotes a alphabetical sequence of lowercase letters (a, b, c *etc.*). The value is based on order of appearance in the article, and is independent of the actual number assigned by the author. Thus the third <compound> is "chem3", no matter what the author has called it. [b] The main sequence of <citgroup> elements has ids of the form "cit1", "cit2", "cit3" *etc.* If the <citgroup> with the id="cit2" is a compound reference containing daughter <citgroups>, these have the form "cit2a", "cit2b", "cit2c" *etc.* [c] Table footnotes should be given an id which is a combination of the table's id and a unique id for the footnote within that table, *e.g.* tab2fna. Table footnotes should be given letters (a, b, c, etc).

Table 3 Mapping of cross-reference element types to target element types

Cross-reference element type	Target element type
<compoundref>	<compound>
<figref>	<figure>
<schemref>	<scheme>
<plateref>	<plate>
<chartref>	<chart>
<eqnref>	<equation>
<boxref>	<box>
<tableref>	<table-entry>
<citref>	<citgroup>
<fnoteref>	<footnote>
<affref>	<aff>

Table 4 Formats for different types of cross-references

Cross-reference type	Type style	Data type
<citref>	superscript	arabic no. [+ letter suffix]
<affref>	superscript to name/address	letter
<fnoteref>	superscript or inline	symbol (*, †, ‡, §, ¶, ‖)
<compoundref>	bold	numbers, letters, roman numerals

However, all of these concepts are allowed to have an id, and some require one — these IDs still need to be specified even if the title or heading itself can be auto-numbered.

5.0 Low-level elements

5.1 Emphasis and font style elements

Changes in font style should be marked up with the appropriate emphasis tags unless they indicate a specific concept, as discussed above under Cross-references.

Individual elements can be used to mark **bold text**, *italic text*, ***bold italic***, SMALL CAPS, ^{superscript} and $_{subscript}$. They can also be used in combination to represent, for example, $^{\mathbf{superscript\ bold\ text}}$.

Several other similar elements are available (underlined text, overbar, underbar) which should **not** be used. If this formatting is found in an article, then a different stategy should be employed; use of combining or special characters (X, δ) for single instances, or ugraphics for complex forms.

Footnotes. All <footnote> elements should be placed just after the first <fnoteref> element that references them.

All footnote characters should be auto-generated. In text, they follow the order:

†, ‡, §, ¶, ‖
then **, ††, ‡‡, §§, ¶¶, ‖‖
then ***, †††, ‡‡‡, §§§, ¶¶¶, ‖‖‖ *etc.*

In table footnotes, they just appear as *a, b, c, d, etc.*, where these letters are taken from the end of the id attribute's value.

Text. The symbols =, +, −, ÷ and × are spaced on either side when in an equation (there is spacing around the mathematical character when it is between two digits *e.g.* 4 + 4. When it is just the character and one digit there is no space *e.g.* +4). This also applies to ∝, ±, ∼, ≈, > and < and their >= variants, and other such symbols.

Multiple <citrefs> shouldn't be spaced:

```
<citref idrefs="cit1 cit4 cit5 cit12">1, 4, 5, 12</citref>
```

should be:

```
<citref idrefs="cit1 cit4 cit5 cit12">1,4,5,12</citref>
```

The <figure>, <scheme>, *etc.* elements should be placed after the end tag of the paragraph in which they are first referenced.

Paragraph (<p>) elements should not be used within <title> tags.

For elements where the content model is empty (<ugraphic>, <colspec>, <icgraphic>) the elements need a closing solidus for XML:

```
<colspec colname="1" colwidth="2.82*" align="left"/>
```

5.2 Tables

Tables will normally appear between paragraphs, marked up according to CALS-compatible XML. The standard CALS attributes should be used to render the table in a form that is as close as possible to the printed result. This includes, but is not limited to, the relative widths of columns, spanning of rows and columns, and the use of lines to separate headings. The specific conventions listed below are intended to be compatible with the approach supported by Arbortext Editor's table editor:

- for relative column widths, use <colspec> with colname="*n*" and colwidth="*X.XX**", where *n* is the column number as an integer, and *X.XX* is proportional to1.00, the default column width;
- individual cells (<entry> element type) refer to their <colspec> with colname="*n*";
- horizontal spans use namest="*n*" nameend="*m*" within <entry>, where *n* and *m* are the column names for the start and end of the span;
- vertical spans use morerows="*x*" within <entry>, where *x* is the number of *extra* rows spanned (*i.e.*morerows="1" gives a cell that spans two rows);
- horizontal alignment within a cell is controlled by the align attribute being set to "center", "right", "justify", "char" or "left" (default) within <entry>. When the align attribute is not specified for <entry>, the value in the appropriate <colspec> element will be used as a fallback;
- vertical alignment within a cell is controlled by the valign attirbute being set to "top", "middle" or "bottom" (default) within <entry>. When the valign attribute is not specified for <entry>, the value in its parent <row> element will be used as a fallback, and failing that the value in the <row> element's own parent (<thead>, <tfoot> or <tbody>);
- do not mark up ruler lines within tables. Default style rules will insert a rule below headings, and at the start and end of the table grid.

However, tables will sometimes be too complex to represent in this way, and so will be prepared as a graphic. To deal with this variation, a 'cover element' <table-entry> is provided, which contains either an inline <table> entry or a <ugraphic>. It is <table-entry> which requires a unique ID for <tableref> elements to point to, and which contains a <title> element.

One side-effect of this approach is that un-numbered tables can simply be encoded as <table>.

5.3 Chemistry

Chemical compounds and simple formulae can often be represented as inline markup. <sup> and <inf> can be used to shift text, and combining entities can be used to place rules above or below chemical symbols. The Unicode character set supports most chemical symbols that will be encountered. The <stack> element type can be used to encode the situation where one character appears directly above another.

Where chemical formulae are too complex to render as inline XML, an inline or displayed <ugraphic> should be used instead.

5.4 Equations

Equations may appear inline, marked up in XML using the tools available.

However, equations will fequently be too complex to represent in this way, and so will be prepared as a graphic. To deal with this variation, a 'cover element' <equation> is provided, which contains either an inline <eqntext> entry or a <ugraphic>. <equation> requires a unique ID for <eqnref> elements to point to.

Multi-line text equations must be set as graphics.

5.5 Citations

Where citations follow the standard pattern for journal articles, the <journalcit> element type should be used. In all other cases (including 'difficult' journal article citations, books, theses, computer software, *etc.*), the more flexible <citation> element type should be used. <citext> should be used to mark up text within the

References section which is not a citation of any kind.

Numbering citations. As noted above in Links and Cross-references, the citation number is a property of the enclosing <citgroup> element, not the citation itself. This makes it easy to deal with the case where more than one citation is given under the same reference number. It also allows running text to be mixed with, or indeed take the place of, proper citations.

Note that the expected pattern for numbering citations is to use numbers for top-level entries, and letters for sub-entries. If the citations follow this pattern, the <no> element should not be provided for any <citgroup> element. Instead, nested <citgroup> elements should be used to represent the lower-level citations. (See the source XML of these instructions for an example of this technique.)

Standard journal citations. Standard journal citations follow this model:

- author (at least one)
- optional article title
- [journal] title
- year
- volume number
- issue number
- first page or page range
- translation (optional)

Unless stated otherwise, each element should appear exactly once, and elements should appear in the order given. In such cases, <journalcit> can and should be used. The citation should be entered as a series of analysed subelements. No punctuation should be recorded between each component of the citation, and no style markup (*e.g.* italic for titles; bold for volume numbers) should be included. Punctuation and styling will be applied by the rendering process. Thus the citation:

> G. H. Jonker and J. H. Van Santen, *Physica,* 1950, **16**, 337

should be encoded:

```
<journalcit><citauth><fname>G. H.</fname><surname>Jonker
</surname></citauth><citauth><fname>J. H.</fname><surname>Van
Santen</surname></citauth><title>Physica</title><year>1950
</year><volumeno>16</volumeno><pages><fpage>337</fpage>
</pages></journalcit>
```

Non-standard citations. The <citation> element type should always be used for non-standard citations which, do not fit the standard model. The type of citation should be specified in the type attribute. Allowed values are:

- article (the default value - this doesn't need to be specified)
- book
- thesis
- patent
- software
- other

Within citations, the following concepts should always be marked up when they are present:

- authors (<citauth>)
- titles (<title>)
- editors (<editor>)
- citation publisher (<citpub>)
- place of publication (<pubplace>)
- year of publication (<year>)
- journal volume number (<volumeno>)
- journal issue number (<issueno>)
- the part of the work that is being cited: section, pagination, *etc.* (<biblscope>)

<citation> elements will be marked up as found, including all punctuation and style changes.

This is an example of a reference to a patent:

> S. Iwaya, H. Masumura, Y. Midori, Y. Oikawa and H. Abe, *US Patent*, 4,404,029, 1983.

This should be encoded:

```
<citation type="patent"><citauth><fname>S.</fname><surname>Iwaya
</surname></citauth>, <citauth><fname>H.</fname><surname>
Masumura</surname></citauth>, <citauth><fname>Y.</fname>
<surname>Midori</surname></citauth>, <citauth><fname>Y.</fname>
<surname>Oikawa</surname></citauth> and <citauth><fname>H.
</fname><surname>Abe</surname></citauth>, <title>US Patent
</title>, 4,404,029, <year>1983</year>.</citation>
```

Book citations. One particular type of non-standard citation which will frequently occur is a reference to a book, either in whole or in part. Again, <citation> should be used to mark these up. The <editor>, <citpub> and <pubplace> element types will often be required within such citations. A fairly typical, simple, example is:

> S. Brooks and B. Johansson, in *Handbook of Magnetic Materials*, ed. K. H. J. Buschow, 1993, 7th edn.

This should be encoded:

```
<citation type="book"><citauth><fname>S.</fname><surname>Brooks
</surname></citauth> and <citauth><fname>B.</fname><surname>
Johansson</surname></citauth>, in <title>Handbook of Magnetic
Materials</title>, ed. <editor>K. H. J. Buschow</editor>,
<year>1993</year>, 7th edn.</citation>
```

Note the following:

- within <citauth>, analysis is the same as for standard citations. No space is required between the forename and surname because the rendering process will add one
- no <it> element is required within the title: it will be rendered as italic
- otherwise, all punctuation (*i.e.* all punctuation between analysed components) is provided exactly as in the source
- the edition information does not fit the model for <biblscope>, and so is left as unanalysed text

A good mixed citation example:

```
<citgroup id="cit5"><citext>During the preparation of this
manuscript, diester <compoundref idrefs="chem1">1</compoundref>
was isolated as a minor side product in the base promoted
rearrangement of the analogous (<it>R</it>′,<it>R</it>′,
<it>R</it>,<it>R</it>)-2,3-butane diacetal (BDA) protected
dimethyl tartrate, see: </citext><journalcit><citauth>
<fname>M. T.</fname><surname>Barros</surname></citauth>
<citauth><fname>A. J.</fname><surname>Burke</surname></citauth>
<citauth><fname>C. D.</fname><surname>Maycock</surname>
</citauth><title>Tetrahedron Lett.</title><year>1999</year>
<volumeno>40</volumeno><pages><fpage>1583</fpage></pages>
</journalcit>
```

and a <citext>:

```
<citgroup id="cit8"><citext>The strong bias towards axial
silylation was seen to fall if the mono sodium alkoxide did
<it>not</it> precipitate prior to addition of the silicon
halide</citext></citgroup>
```

RSC journal abbreviations. The journals published by the RSC have two letter abbreviations (Table 6), which can be used within the XML framework, *e.g.* in <journalref> elements.

Table 6 Journal names and codes

AN	The Analyst
CC	Chemical Communications
CE	CrystEngComm
CP	Physical Chemistry Chemical Physics
CS	Chemical Society Reviews
DT	Dalton Transactions
EM	Journal of Environmental Monitoring
FD	Faraday Discussions
GC	Green Chemistry
IC	Annual Reports (Inorganic)
JA	JAAS
JM	Journal of Materials Chemistry
LC	Lab on a Chip
MB	Molecular BioSystems
NJ	New Journal of Chemistry
NP	Natural Product Reports
OB	Organic & Biomolecular Chemistry
OC	Annual Reports (Organic)
PC	Annual Reports (Physical)
PP	Photochemical & Photobiological Sciences
SM	Soft Matter

5.6 Lists

Lists can be entered as a <list>, containing an optional <head> and any number of <item> elements. Note that, since <list> can occur within <item>, it is possible to declare lists nested to any depth.

5.7 General

If there are internal references that are in effect impossible, just put the text in and leave out the reference. It would helpful to advise us in case an amendment to the schema may be wise, but usually these are one-offs. One case recently had a number of equations in a single ugraphic, itself called Scheme 1. In this case it was not possible to add eqnrefs to the scheme.

A.0 Appendix A. Alphabetical list of element types

A.1 Element definitions

This section contains a definition of every element type in the RSC document type, including element types which are not required for the data capture work. These additional element types are included for editorial use within RSC, or to support future processing of the encoded articles. They are indicated with an asterisk (*).

- <a>*
- <above>. The top half of a <stack>. Contains 'characters only'.
- <abstract>. An abstract of the article. Contains 'text or paragraphs'.
- <ack>. Acknowledgements for the article. Contains 'text or paragraphs'. *Attributes*:
  - title**:** an optional non-standard title for the acknowledgements section.
- <address>. A complete postal address made of a sequence of address subelements (<city>, <postcode>, <state>, <country>, <addrelt>) each separated by spacing but no punctuation.
- <addrelt>. An element within a postal address. Used only when no more specific element type (*e.g.* <city>) is appropriate. Can contain 'simple text'.
- <admin-event>*
- <advert>*
- <aff>. An author's affiliation. Contains one or more pairs of <org> (optional) and <address> (mandatory) followed by any contact details which apply (<phone>, <fax>, <email>, <url>). *Attributes*:
  - id**:** a unique identifier for this affiliation element. See above for guidance on assigning ids.
- <affref>*
- <agent>*
- <annotation>*
- <annotationgrp>*
- <annref>*
- <appendix>. An appendix to an article. Contains an optional <no> and one or more <section>s. *Attributes*:
  - id**:** a unique identifier for this appendix element. See above for guidance on assigning ids.
- <appmat>. A container for appendix matter. See above for general guidance. Contains one or more <appendix> elements.
- <art-admin>. A container for administrative information relating to an article. Contains, in the order specified: <ms-id> (required), <doi> (optional), <pii> (optional), <sici> (optional), <office> (optional), <received> (optional and repeatable), <date> (optional and repeatable), <admin-event> (optional and repeatable).
- <art-back>. A container for an article's back matter. Contains, in the order specified: <ack> (optional), <biblist> (required), <compoundgrp> (required), <section> (optional and repeatable).
- <art-body>. A container for an article's body matter. See above for general guidance. Contains one or more <section>s, or one or more <news-section>s.
- <art-front>. A container for an article's front matter. See above for general guidance on analysing front matter. Contains a <link>, or the following elements in the order specified: <titlegrp> (required), <authgrp> (optional), <conference> (optional), <art-toc-entry> (optional), <arttoc> (optional), <dedicate> (optional), <biography> (optional and repeatable), <abstract> (optional and repeatable), <subject> (optional and repeatable), <keyword> (optional and repeatable).
- <art-links>. A container for links from an article to other resources. Contains any number of <suppinf> and/or <fulltext> elements.
- <art-toc-entry>. Container for resources to use when creating the article's entry in the table of contents for a journal issue. Contains, in the following order: <ictext>, <icgraphic>.
- <article>. An article. Contains a <link> element, or the following elements in the order specified: <art-admin> (optional), <published> (optional and repeatable), <art-links> (optional), <art-front> (optional), <art-body> (optional), <appmat> (optional), <art-back> (optional). *Attributes*:
  - dtd**:** a FIXED attribute which specifies which version of the schema was in use when this XML document was created. There is no need to enter a value for this attribute (and any value other than "RSCPA*x.y*" for version *x.y* of the schema will render the whole article invalid)
  - type**:** the class of article, *e.g.* 'feature', 'communication'. The article type should be taken from the list of codes given above, *e.g.*"ART" for a Paper
- <articleref>*
- <arttitle>. An article title within a citation or journalcit. Contains 'simple text or paragraphs'.
- <arttoc>*
- <authgrp>. A container for details of authors and their affiliations. Contains one or more <author> elements, followed by one or more <aff>s.
- <author>. One author of an article. Repeat for each distinct author. Contains a <person>, followed by an optional <footnote>. *Attributes*:
  - aff**:** one or more idrefs (separated by spaces), specifying which <aff> elements apply to this author
  - role**:** can take the value "corres" (corresponding author)
- <below>. The bottom half of a <stack>. Contains 'characters only'.
- <bi>. Indicates that the contained text should be rendered as bold italic. This is preferable to using separate <bo> and <it> elements. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.
- <biblist>. A container for the bibliography at the end of an

article. Contains a mixture of text and citgroups. *Attributes*:

- title: a non-standard title for the bibliography. Can include a section number, if one is required.

- <biblscope>. The scope of a citation within the work cited. Can include references to sections, chapters, page ranges, *etc.* Contains 'simple text'.
- <biography>. A person's biography. Contains a <link>, or one or more sections.
- <bo>. Indicates that the contained text should be rendered as bold. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type (specifically <compoundref>, which is the most common reason for bold-face within article text).
- <board>*
- <book-review>. A book review, consisting of the citation of the book being reviewed, reviewer's details, and the review itself. Contains a <citation>, followed by an optional <authgrp> for the reviewer's details (*i.e.* the 'author' of the review), followed by one or more paragraphs (<p>) and/or 'inter-paragraph elements'.
- <box>. a floating text box. Contains a single <section>. *Attributes*:
  - id: a mandatory unique identifier for this element
  - pos: can optionally take the value "fixed" to indicate that the <box> cannot float

- <boxref>. A reference to a floating text box. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs: one or more space-separated idref's, specifying the box(es) to which cross-reference is being made

- <byline>*
- <chart>. A chart. Contains an optional <title>. See above for general guidance on encoding graphics. *Attributes*:
  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc: the filename of the graphic.

- <chartref>. A cross-reference to a chart. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs: one or more space-separated idrefs, specifying the chart(s) to which cross-reference is being made

- <citation>. Container for an individual citation that doesn't fit the model for a standard journal citation (<journalcit>). Should only be used if <journalcit> cannot. See above for general guidance on encoding citations. Contains mixed content, which can include the following element types as required: <citauth>, <title>, <year>, <volumeno>, <issueno>, <arttitle>, <biblscope>, <editor>, <citpub>, <pubplace>, <link>, <url>, <email> and <trans> 'emphasis' elements. *Attributes*:
  - type: the type of citation.

- <citauth>. An author within a <citation>/<journalcit> element. Contains an optional <fname> followed by a mandatory surname.
- <citext>. Citation text. Used to encode material found within a citations list that is not part of a <journalcit> or <citation>. Contains 'simple text'.
- <citgroup>. A group of citations with a single reference number. (Most <citgroup>s will only contain a single <journalcit> or <citation> element.) See above for general guidance on encoding citations. Contains one or more of the following, in any order: <citext>, <journalcit>, <citation>, <citgroup>. A <commentary> element may also appear after the various elements above. *Attributes*:
  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.

- <citpub>. The publisher of a citation. Contains 'simple text'.
- <citref>. A reference to a citation. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs: one or more space-separated idrefs, specifying the citation(s) to which cross-reference is being made.
  - position: can take the value "super" or "baseline" ("super" is assumed if the attribute is not set).

- <city>. The name of a city. Must consist of character data only.
- <coden>*
- <colspec>. A specification of the characteristics of a column in a table. Empty element: has no data content. *Attributes*:
  - colname: the column's name (usually the column number)
  - colwidth: the column's width, as a relative fraction of 1.00 (= average column width given equal spacing)
  - align: the alignment of the column's content
  - char: the character to be used for alignment within the column
  - charoff: the offset for character alignment within the column

- <commentary>. A description of the value of a citgroup. Contains 'simple text'. *Attributes*:
  - rating: used to denote the rank of the citation: 0 (default), 1 or 2

- <compname>*
- <compound>. Specifies the id of a chemical compound. Optionally contains one or more <link> or <namedent-xref> elements. *Attributes*:
  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.

- <compoundgrp>. A container for zero or more <compound> and <compoundset> elements. A <compoundgrp> is required at the end of each article so that <compoundref> elements have a target to point to.
- <compoundref>. A reference to a chemical compound. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs: a single idref, specifying the <compound> or <compoundset> to which cross-reference is being made.

- <conference>*
- <confgrp>*
- <confname>*
- <contact>*
- <country>. A country name. Must consist of character data only.
- <cpyrt>*
- <data>*
- <datagrp>*
- <dataref>*
- <date>. A general year-month-day date. Contains a <year>,

followed by an optional <month> and an optional <day>. *Attributes*:

- role: the role played by this date (*e.g.*"accepted" or "revised")

- <daterange>*
- <day>. A numerical day: 1/2/3/.../31. Should not contain anything apart from the day number itself.
- <dd>. A definition description, part of a <deflist>. Contains 'text or paragraphs'.
- <dedicate>. A dedication. Contains 'text or paragraphs'.
- <def>. The definition of a term, part of a <deflist>. Contains the <term> itself, followed by its definition in a <dd>.
- <deflist>. A definition list, containing an optional <head>, and one or more definitions <def>).
- <denom>. The denominator of a fraction. Contains 'simple text'.
- <doi>. A Digital Object Identifier. Contains character data only.
- <editnote>*
- <editor>. The editor of an article or book. Contains 'simple text'.
- <email>. An e-mail address. Contains character data only. Only enter the actual address: the prefix "E-mail:" will be generated by style sheets.
- <entry>. An entry (cell) in a table. See above for general guidance on encoding tables. Contains mixed content which can include text elements, graphics, and equations. *Attributes*:

  - colname: the name of the column in which this cell appears
  - namest: the name of the start column for this cell
  - nameend: the name of the end column for this cell
  - morerows: the number of rows occupied by this cell
  - align: the alignment of the column's content
  - char: the character to be used for alignment within the column
  - valign: the vertical alignment of the column's content

- <eqnref>. A reference to an equation. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:

  - idrefs: one or more space-separated idrefs, specifying the equation(s) to which cross-reference is being made

- <eqntext>. An equation expressed in textual form. See above for general guidance on encoding equations. Contains 'simple text or paragraphs'. Use <p>s to lay out multi-line equations. *Attributes*:

  - display: can take the value "displayed" (default) or "inline". Use this attribute to indicate whether the equation should be set as a separate block, or rendered inline.

- <equation>. An equation. See above for general guidance on encoding equations. Contains an optional <no>, followed by a textual equation (<eqntext>) or a graphic displaying the equation (<ugraphic>). *Attributes*:

  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.

- <extnl-data>*
- <extnl-dataref>*
- <fax>. A fax number. Can only contain character data.
- <figref>. A cross-reference to a figure. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:

  - idrefs: one or more space-separated idrefs, specifying the figure(s) to which cross-reference is being made

- <figure>. A figure. Contains an optional <title>. See above for general guidance on encoding graphics. *Attributes*:

  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc: the filename of the graphic.

- <fname>. A person's first name. Contains 'simple text'.
- <fnoteref>. A reference to a footnote (at the end of the article, or in the footer of a table). Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on footnotes, and on creating cross-references. *Attributes*:

  - idrefs: a **single** idref, specifying the footnote to which cross-reference is being made

- <footer>. A sequence of paragraphs at the end of a news item, typically set in italic. Contains one or more <p>s.
- <footnote>. A footnote in the article, or in a table footer. See above for general guidance on encoding footnotes. Footnotes in the article are placed at the point where the first footnote reference is to appear in the rendered result. Contains text or paragraphs. *Attributes*:

  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.

- <fpage>. The number of the first page within an issue on which the printed version of an article appears. Can only contain character data.
- <fraction>*
- <fulltext>*
- <group>*
- <head>. A heading (*e.g.* for a list, index, or definition list). Contains paragraphs or text.
- <icgraphic>. A graphic to be included in an illustrated contents list entry. Empty element: has no contents. See above for general guidance on encoding graphics. *Attributes*:

  - id: a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc: the filename of the graphic.

- <ictext>. Text describing the article, to be included in an illustrated contents list entry. Contains paragraphs or text.
- <index>*
- <index-entry>*
- <inf>. Inferior (subscript) text. Indicates that the contained text should be rendered as subscript. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.
- <info>*
- <issn>*
- <issue>*
- <issue-back>*
- <issue-front>*
- <issue-toc>*
- <issueid>*
- <issueno>. The issue number within a volume. Can only contain character data.
- <issueref>*
- <it>. Indicates that the contained text should be rendered as italic. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.

- <item>. An item within a list. See above for general guidance on encoding lists. Contains paragraphs or 'simple text'.
- <jnltrans>. A translation of a simple journal citation (<journalcit>). Also used for Chem. Abstracts references, with the abstract number in <fpage>. Contains the following, in the order specified: <sertitle> (optional), <year> (optional), <volumeno> (optional), <pages> (optional and repeatable).
- <journal>*
- <journalcit>. A citation which follows the standard model for simple citations of journal articles. Use <citation> for more complex cases, and for citations to anything other than journal articles. Use <citext> only for text within the References section which is not a citation at all. See above for general guidance on encoding citations. Contains these elements in the order specified: <citauth> (one or more), <title>, <year>, <volumeno> (optional), <issueno> (optional), <pages>, <jnltrans> (optional), <link> (optional and repeatable).
- <journalref>*
- <keyword>*
- <link>. A link to [part of] another document or resource. Contains simple text. This content will be a unique identifier for the resource, *e.g.* a journal code or an article's manuscript number. *Attributes*:
  - type**:** the type of link, *e.g.*"DOI" for DOI cross-references
- <list>. A list. See above for general guidance on encoding lists. Contains an optional head, followed by one or more <item>s.
- <location>*
- <logo>*
- <lpage>. The number of a printed article's last page. Contains character data only.
- <member>*
- <month>. A month. Contains character data only. Months should be specified in full, *e.g.* 'January'.
- <ms-id>. The RSC's unique identifier for an article. Contains character data only. Conventions for formatting article identifiers are given above. **To be added by data capture agency**
- <nameelt>. A component of an organisation's name. Contains 'simple text'.
- <news-article>. A full article (with title and author details, and back matter such as a list of citations) found within a news section. Contains these elements, in the order specified: <art-front>, <art-body>, <appmat>, <art-back>.
- <news-item>. A relatively simple news item. For more complex material, use <news-article> instead. Contains these elements, in the order specified: <title> (optional), <authgrp> (optional), <abstract> (optional), <p> or paragraph-level elements (optional and repeatable), <footer> (optional).
- <news-section>. A container for one or more news articles or (more usually) news items, plus other formats such as advertisements and conference listings. Can contain nested <news-section>s to support *e.g.* a two-level structure of news sections. Contains an optional <title>, followed by zero or more of the following, in any order: <news-section>, <news-article>, <news-item>, <book-review>, <advert>, <info>, <confgrp>, <p>, paragraph-level elements.
- <no>. A number or other identifier (for a table, figure, *etc.*). Contains character data only. See above for general guidance on numbering strategy.
- <no-of-pages>. The number of pages in the printed version of an article. Contains character data only.
- <note>. A note. Contains text or paragraphs.
- <numer>. The numerator of a fraction. Contains 'simple text'.
- <office>. The RSC office responsible for managing an article. Contains character data only.
- <org>. An organisation's name and address. Contains a <link>, or one or more <orgname>s followed by zero or more <address> tags.
- <orgname>. An organization's name. Contains one or more <nameelt>s.
- <overbar>*
- <p>. A paragraph. Contains mixed content (*i.e.* text and subelements intermixed).
- <pages>. The range of pages covered by a citation. Contains a <fpage>, optionally followed by a <lpage>.
- <persname>. A person's name. Contains the following, in the order specified: <qualifier> (optional), <fname> (optional), <surname>, <qualifier> (optional).
- <person>. Details about a person. Contains a <link>, or the following elements in the order specified: <persname> (required; repeatable), <biography> (optional), <address> (optional and repeatable).
- <phone>. A telephone number. Contains character data only.
- <pii>*
- <plate>. A plate. Contains an optional <title>. See above for general guidance on encoding graphics. *Attributes*:
  - id**:** a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc**:** the filename of the graphic.
- <plateref>. A reference to a plate. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs**:** one or more space-separated idrefs, specifying the plate(s) to which cross-reference is being made
- <postcode>. A postcode. Contains character data only.
- <pubfront>. Publication front matter. Contains the following elements in the order specified: <fpage>, <lpage> (optional), <no-of-pages>, <date>.
- <published>. A link to a document/resource in which an article has been published. Contains a <citext>, or the following elements in the order specified: <journalref>, <volumeref> (optional), <issueref> (optional), <pubfront> (optional). *Attributes*:
  - type**:** the type of publication. Should take one of the values: "print", "web" or "subsyear".
- <publisher>*
- <pubplace>. The place of publication of a book, *etc.* Contains 'simple text'.
- <qualifier>. A qualification to a person's name, such as a title, an honorific, or a phrase such as 'the late'. Contains 'simple text'.
- <quotation>. Indicates that the contained text is a quotation from a previous work.
- <received>. A container for details of the date when, and place where, an article was received. Contains an optional <city>, followed by a <date>.
- <resource>*
- <resourcegrp>*
- <resourceref>*
- <role>*
- <roman>. Indicates that the contained text should be rendered as a roman typeface. Contains 'simple text'. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.
- <row>. A row in a table or table heading. See above for general guidance on encoding tables. Contains one or more <entry> elements.
- <sansserif>. Indicates that the contained text should be rendered in a sans serif typeface. Contains 'simple text'. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.

- <scheme>. A scheme. Contains an optional <title>. See <above> for general guidance on encoding graphics. *Attributes*:
  - id**:** a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc**:** the filename of the graphic.

- <schemref>. A reference to a scheme. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs**:** one or more space-separated idrefs, specifying the plate(s) to which cross-reference is being made

- <scp>. Indicates that the contained text should be rendered in small caps. Contains 'simple text'. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type.
- <section>. A top-level section. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect1> (optional and repeatable).
- <sercode>*****
- <sertitle>*****
- <sici>*****
- <stack>. One or more characters appearing directly above other characters (like a fraction without the horizontal line). Contains <above> followed by <below>.
- <state>. A geopolitical unit such as a state, county, *etc.* Contains character data only.
- <subject>*****
- <subsect1>. A level-1 subsection. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect2> (optional and repeatable).
- <subsect2>. A level-2 subsection. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect3> (optional and repeatable).
- <subsect3>. A level-3 subsection. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect4> (optional and repeatable).
- <subsect4>. A level-4 subsection. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect5> (optional and repeatable).
- <subsect5>. A level-5 subsection. Contains these elements in the order specified: <no. (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable), <subsect6> (optional and repeatable).
- <subsect6>. A level-6 subsection. Contains these elements in the order specified: <no> (optional), <title> (optional), <p> or paragraph-level elements (optional and repeatable), <deflist> (optional and repeatable).
- <subtitle>. A [table] subtitle. Contains 'simple text' or paragraphs.
- <sup>. Indicates that the contained text should be rendered in superscript. Contains 'simple text'. Only use this element when it is not possible to deduce *why* the text is rendered in this way. If possible, always use a more meaningful element type. <sup> is often mistakenly used instead of <citref>.
- <suppinf>. Contains a <link> to supplementary information for an article.
- <surname>. A surname. Contains 'simple text'.
- <table>. A table, encoded using CALS-compliant XML markup. See above for general guidance on encoding tables. (Tables which cannot be thus encoded should be prepared as images, and encoded as <ugraphic>s.) Contains an optional <title>, followed by an optional <subtitle>, followed by one or more <tgroup>s. Note that <title> and <subtitle> within <table-entry> should be used in preference to these elements, since this allows titles for XML-encoded and 'image' tables to be treated consistently.
- <table-entry>. 'cover group' for a table, whether declared inline as <table> or given as a <ugraphic>. See above for general guidance on encoding tables. Contains an optional <title>, followed by an optional <subtitle>, followed by either <table> or <ugraphic>. *Attributes*:
  - id**:** a mandatory unique identifier for this element. See above for guidance on assigning ids.

- <tableref>. A reference to a table. Contains 'emphasised text' giving a human-readable description of the cross-reference. See above for general guidance on creating cross-references. *Attributes*:
  - idrefs**:** one or more space-separated idrefs, specifying the plate(s) to which cross-reference is being made

- <tbody>. A table's body matter (*i.e.* the main table, ignoring any header or footer). See above for general guidance on encoding tables. Contains one or more <row>s.
- <term>. A term being defined in a <deflist>. Contains 'simple text'.
- <textref>*****
- <tfoot>. The footer area of a table. See above for general guidance on encoding tables. Contains zero or more <colspec>s, followed by one or more <row>s.
- <tgroup>. A table group. See above for general guidance on encoding tables. Contains these elements, in the order specified: <colspec> (optional and repeatable), <thead> (optional), <tfoot> (optional), <tbody>. *Attributes*:
  - cols**:** the number of columns in the table

- <thead>. The header area of a table. See above for general guidance on encoding tables. Contains zero or more <colspec>s, followed by one or more <row>s.
- <title>. A title (of a figure, table, journal, *etc.*). Contains 'simple text' or paragraphs.
- <titlegrp>. A container for an article's main titles. Contains one or more <title>s.
- <toc-entry>*****
- <toc-head>*****
- <trans>. A translation (of a citation). Contains mixed content, which can include the following element types as required: <citauth>, <title>, <year>, <volumeno>, <issueno>, <arttitle>, <biblscope>, <editor>, <citpub>, <pubplace>, <link>, <url>, <email>, <trans>, 'emphasis' elements.
- <ugraphic>. An untitled graphic. Use this element to encode any graphical content which doesn't have a title. See above for general guidance on encoding graphics. *Attributes*:
  - id**:** a mandatory unique identifier for this element. See above for guidance on assigning ids.
  - xsrc**:** the filename of the graphic.
  - display**:** how the graphic is to be displayed. Set as "inline" or leave blank.

- *****
- <underbar>*****
- <unknown>*****
- <url>. A URL. Contains character data only.
- <value>*****

- <volume>*
- <volumeno>. A journal volume number. Contains character data only.
- <volumeref>*
- <warning>*
- <who>*
- <year>. A four-digit year. Contains character data only. The value "PENDING" is allowed for <date> within <pubfront>.

References

1 http://www.oasis-open.org/html/a502.htm
2 http://www.oasis-open.org/html/a503.htm